Value of crowd-based water level class observations for hydrological model calibration

Etter, Simon ; Strobl, Barbara ; Seibert, Jan ; van Meerveld, H J

Abstract: While hydrological models generally rely on continuous streamflow data for calibration, previous studies have shown that a few measurements can be sufficient to constrain model parameters. Other studies have shown that continuous water level or water level class (WL-class) data can be informative for model calibration. In this study, we combined these approaches and explored the potential value of a limited number of WL-class observations for calibration of a bucket-type runoff model (HBV) for four catchments in Switzerland. We generated synthetic data to represent citizen science data and examined the effects of the temporal resolution of the observations, the numbers of WL-classes, and the magnitude of the errors in the WL-class data on the model validation performance. Our results indicate that on average one observation per week for a one-year period can significantly improve model performance compared to the situation without any streamflow data. Furthermore, the validation performance for model parameters calibrated with WL-class observations was similar to the performance of the calibration with precise water level measurements. The number of WL-classes did not influence the validation performance noticeably when at least four WL-classes were used. The impact of typical errors for citizen-science-based estimates of WL-classes on the model performance was small. These results are encouraging for citizen science projects where citizens observe water levels for otherwise ungauged streams using virtual or physical staff gauges.

DOI: https://doi.org/10.1029/2019wr026108

Value of crowd-based water level class observations for hydrological model calibration

S. Etter[1], B. Strobl[1], J. Seibert[1,2] and H.J. van Meerveld[1]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.

[2]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, P.O. Box 7050, 75007 Uppsala, Sweden.

Corresponding author: Simon Etter (simon.etter@geo.uzh.ch)

Key Points

- Water level class observations can be informative for hydrological model calibration.
- Model parameters calibrated with water level class data performed similarly well as those calibrated with precise water level measurements.
- Errors in water level class data observations had a minimal effect on the streamflow simulations

Abstract

While hydrological models generally rely on continuous streamflow data for calibration, previous studies have shown that a few measurements can be sufficient to constrain model parameters. Other studies have shown that continuous water level or water level class (WL-class) data can be informative for model calibration. In this study, we combined these approaches and explored the potential value of a limited number of WL-class observations for calibration of a bucket-type runoff model (HBV) for four catchments in Switzerland. We generated synthetic data to represent citizen science data and examined the effects of the temporal resolution of the observations, the numbers of WL-classes, and the magnitude of the errors in the WL-class data on the model validation performance. Our results indicate that on average one observation per

week for a one-year period can significantly improve model performance compared to the situation without any streamflow data. Furthermore, the validation performance for model parameters calibrated with WL-class observations was similar to the performance of the calibration with precise water level measurements. The number of WL-classes did not influence the validation performance noticeably when at least four WL-classes were used. The impact of typical errors for citizen-science-based estimates of WL-classes on the model performance was small. These results are encouraging for citizen science projects where citizens observe water levels for otherwise ungauged streams using virtual or physical staff gauges.

Plain Language Summary

Normally, multiple years of streamflow measurements are used to calibrate a hydrological model for a specific catchment so that it can be used to, for instance, predict floods or droughts. Taking these measurements is expensive and requires a lot of effort. Therefore, such data are often missing, especially in remote areas and developing countries. We investigated the potential value of water level class data for model calibration. Water level classes can be observed by citizens with the help of a virtual ruler with different classes that is pasted onto a picture of a stream shore as a sticker.

We show that one WL-class observation per week for one year improves model calibration compared to situations without streamflow data. The model results for the WL-class observations were as good as precise water level observations that require a physical staff gauge or continuous water level data measurements that can be obtained from a water level sensor that is installed in the stream. However, the results were not as good as when streamflow data were used for model calibration, but these are more expensive to collect. Errors in the WL-class observations did in most cases not affect the model performance noticeably.

1 Introduction

Hydrological models are usually calibrated with continuous streamflow data acquired at gauging stations. Such datasets are scarce, especially for remote regions and developing countries, even though people in these areas are often affected by various kinds of water issues (Mulligan, 2013). Globally, hydrological observation networks are on the decline, mainly due to reduced financial resources (Kundzewicz, 1997). Furthermore, access to available data is often restricted (Fekete et al., 2012). To collect data in ungauged basins, citizen science approaches that use modern communication technology (i.e., smartphones) can be helpful. Citizen science approaches can also incorporate local knowledge, for instance, for hazard assessment (Sy et al., 2018) and help to raise public awareness of environmental issues (Lanfranchi et al., 2014). However, the usefulness of citizen science data is often questioned due to, for example, the perceived lack of experience of the volunteers (Cohn, 2008) and potential biases, such as location bias related to the population density or temporal bias related to the timing of the observations (Kosmala et al., 2016). It is important to standardize measurement protocols (Dickinson et al., 2012), e.g., by using smartphone applications, to evaluate the accuracy and value of the collected data, and to improve the measurement protocols iteratively when needed. It is also useful to thoroughly examine the potential use of citizen science data before starting a new project.

Publications that include citizen science projects focusing on water quantity in streams are still rather scarce; most publications on water related citizen science projects have focused on water quality (Buytaert et al., 2014; Njue et al., 2019). Some recent examples of water quantity-focused projects are the EU-funded citizen observatories that aim to complement data collection by authorities, such as WeSenseIt (www.wesenseit.com; Lanfranchi et al., 2014), GroundTruth2.0 (https://gt20.eu) and SCENT (https://scent-project.eu). Projects that specifically focus on streamflow or water levels are: CrowdHydrology in the US (Lowry et al., 2019; Lowry & Fienen, 2013), Smartphones4Water in Nepal (www.smartphones4water.org; Davids et al., 2017), a project in Kenya (www.uni-giessen.de/hydro/hydrocrowd_kenya; Weeser et al., 2018), Cithyd in Italy (www.cithyd.com; Balbo & Galimberti, 2016), and CrowdWater (www.crowdwater.ch; Seibert et al., 2019). The CrowdWater project aims to explore the value of citizen science data and to collect water level class (WL-class) data (Seibert et al., 2019), as well as qualitative data on soil moisture and the state of temporary streams (Kampf et al., 2018;

Seibert et al., 2019), and riverine export of macro plastic. For observations of WL-classes, virtual staff gauges with class markings are inserted onto a photograph of the streambank, bridge pillar, or other features in the stream. These features and the virtual staff gauge then serve as a reference to which later observations of the water level are compared. Repeated observations result in time series of WL-classes. However, these series are irregular in time and potentially contain observation errors (Strobl et al., 2019a).

Several studies have examined the value of discontinuous streamflow data for the calibration of hydrological models. For example, Pool et al. (2019) investigated the value of a limited number of streamflow measurements for calibration of the HBV model (Bergström, 1976; Lindström et al., 1997) and found that twelve measurements taken during a one-year period can lead to satisfying model simulations. Seibert & McDonnell (2015) showed for the Maimai catchment in New Zealand that streamflow measurements throughout an event or ten observations during high flow periods provide as much information for model calibration as three months of continuous measurements. These model studies assumed error-free streamflow measurements. All measurements are affected by errors, and these can be considerable for streamflow measurements (particularly during high flows or low flows; McMillan et al., 2018), but for citizen science data, errors might be particularly large (Aceves-Bueno et al., 2017). This can significantly limit the value of the data. Therefore, we previously investigated the value of streamflow data that included errors that are typical for citizen based estimates of streamflow (Etter et al., 2018). We found that streamflow estimates from citizens, who did not receive any form of training, did not improve model performance compared to a model with random parameter sets. We concluded that either the errors in the streamflow estimates have to be reduced by some form of training, or that a quantity, that is easier to estimate, such as water levels or WL-classes, should be used (Strobl et al., 2019a). Water level measurements require the installation of a staff gauge. Citizens then can read the water level from the staff gauge and report them via text messages or a mobile application. Previous studies have shown that this method works well and can provide useful and accurate data (Lowry et al., 2019; Weeser et al., 2018). However, the installation of a staff gauge can be complicated in practice. Beyond issues such as how to securely fix the gauge, permissions by local authorities might be required. Obtaining permits can require time and effort and cause additional costs. WL-class estimates, as used within the CrowdWater project, do not require a physical staff gauge and are, thus, more

scalable. However, the data have a lower precision (and likely also lower accuracy) than readings from a staff gauge.

Continuous (e.g., daily) water level or WL-class data can be informative for hydrological model calibration. Seibert & Vis (2016) concluded that the use of daily water level data for model calibration results in a surprisingly good model performance, especially for humid catchments. For arid regions additional information was necessary to achieve a good simulation. In another study, van Meerveld et al. (2017) showed that daily WL-class data are informative for hydrological model calibration as well, and that the performance of the model calibrated with WL-class data with at least five equally frequent classes, was not much worse than a model calibrated with water level data.

In this study, we used a similar approach as in Etter et al. (2018) in order to be able to compare the results. We aim to develop a methodology that is quick and easy to use for citizen scientists, while at the same time being robust and informative for the calibration of hydrological models. We therefore investigated the potential value of discontinuous WL-class data as these can be obtained by citizens using synthetic data. Our objectives were to (i) assess the potential value of a few WL-class observations at intervals that are realistic for citizen science projects, for model calibration, (ii) assess the potential effect of likely errors in WL-class observations on model performance, and (iii) investigate the influence of the number of WL-classes in combination with different observation scenarios on model performance.

2 Methods

At the time of writing this paper, an insufficient number of repeated observations had been collected with the CrowdWater App to determine the value of WL-class data for model calibration. We, therefore, used synthetic data (cf. Etter et al., 2018; van Meerveld et al., 2017; Seibert & Vis, 2016), which is an efficient approach to assess data-requirements before making considerable efforts to collect the data (Christophersen et al., 1993; Pool et al., 2019). First, we converted the water level time series for four Swiss catchments into WL-class time series. From these continuous datasets, we created time series with fewer data points representing different observation scenarios and introduced errors that are typical for citizen estimates of WL-classes

(Strobl et al., 2019a). We then used these synthetic datasets to calibrate a simple bucket-type model, the HBV model (Bergström, 1976; Lindström et al., 1997; Seibert & Vis, 2012). Finally, we used the calibrated parameter sets to evaluate the model performance for the validation period by comparing it to the observed streamflow. We compared the validation performance to the validation performance of the model calibrated with the original (continuous, and assumed to be error free) streamflow data (upper benchmark), and the validation performance of the non-informed case, where the model is run with random parameter sets (lower benchmark).

2.1 Catchments

For this study, we selected four gauged catchments in Switzerland with flow regimes (Aschwanden & Weingartner, 1985). Streamflow measurements at the outlet of these catchments have good quality for both high and low flow conditions and are unaffected by backwater issues. Furthermore, the catchments are relatively little affected by anthropogenic influences and have no glaciers. The catchment areas range from 79 to 186 km^2 and the mean elevations from 652 to 1651 m a.s.l. (Table 1 and Figure 1).

2.2 HBV model

We used the bucket-type hydrological model HBV (Lindström et al., 1997), which was originally developed at the Swedish Meteorological and Hydrological Institute (SMHI) by Bergström (1976). The HBV model consists of routines for snow storage, soil water and groundwater. In this study, we used the model implementation HBV-light (Seibert & Vis, 2012). The catchments were divided into elevation zones, each covering a band of 100 m, for which the snow, soil, and groundwater routines were computed individually.

Table 1 Catchment characteristics for the four Swiss catchments used in this study. Long-term annual averages were computed for the period 1974-2014, except for Verzasca for which the 1990-2014 period was used.

Catchment		Murg	Guerbe	Mentue	Verzasca
Gauging station (FOEN station number)		Waengi (2126)	Belp, Mülimatt (2159)	Yvonand, La Mauguettaz (2369)	Lavertezzo, Campiòi (2605)
Weather stations		Aadorf-Taenikon, Hörnli	Plaffeien, Bern-Zollikofen	Mathod, Pully	Acquarossa, Cimetta, Magadino, Piotta
Area [km^2]		79	117	105	186
Elevation	Min	465	522	445	490
[m a.s.l.]	Max	1035	2176	927	2864
Regime Type[1]		Pluvial-inférieur	Pluvial-supérieur	Pluvial-jurassien	Nivo-pluvial-méridional
Min / Max Pardé coefficients		0.68 / 1.34	0.77 / 1.39	0.46 / 1.57	0.23 / 2.22
Mean annual streamflow Q [mm/y]		756	746	491	1764
Mean annual precipitation P [mm/y]		1343	1319	1287	2014
Mean runoff ratio (Q/P)		0.56	0.57	0.38	0.88
July – September streamflow [mm] (calibration | validation)					
Dry		90 | 86	106 | 94	26 | 24	324 | 307
Average		125 | 149	202 | 195	54 | 62	417 | 439
Wet		220 | 228	308 | 451	93 | 187	670 | 810
Annual runoff ratio (calibration | validation)					
Dry		0.72 | 0.54	0.37 | 0.82	0.41 | 0.41	0.98[2] | 0.71
Average		0.55 | 0.43	0.48 | 0.60	0.52 | 0.65	0.66 | 0.63
Wet		0.56 | 0.54	0.54 | 0.81	0.50 | 0.52	1.32[2]| 0.73

[1] *Regime types according to Aschwanden & Weingartner (1985)*

[2] *For Verzasca the calibration years 2011 and 2013 have an unrealistic runoff-rainfall ratio (>0.9) and were therefore excluded from all simulations (see text).*

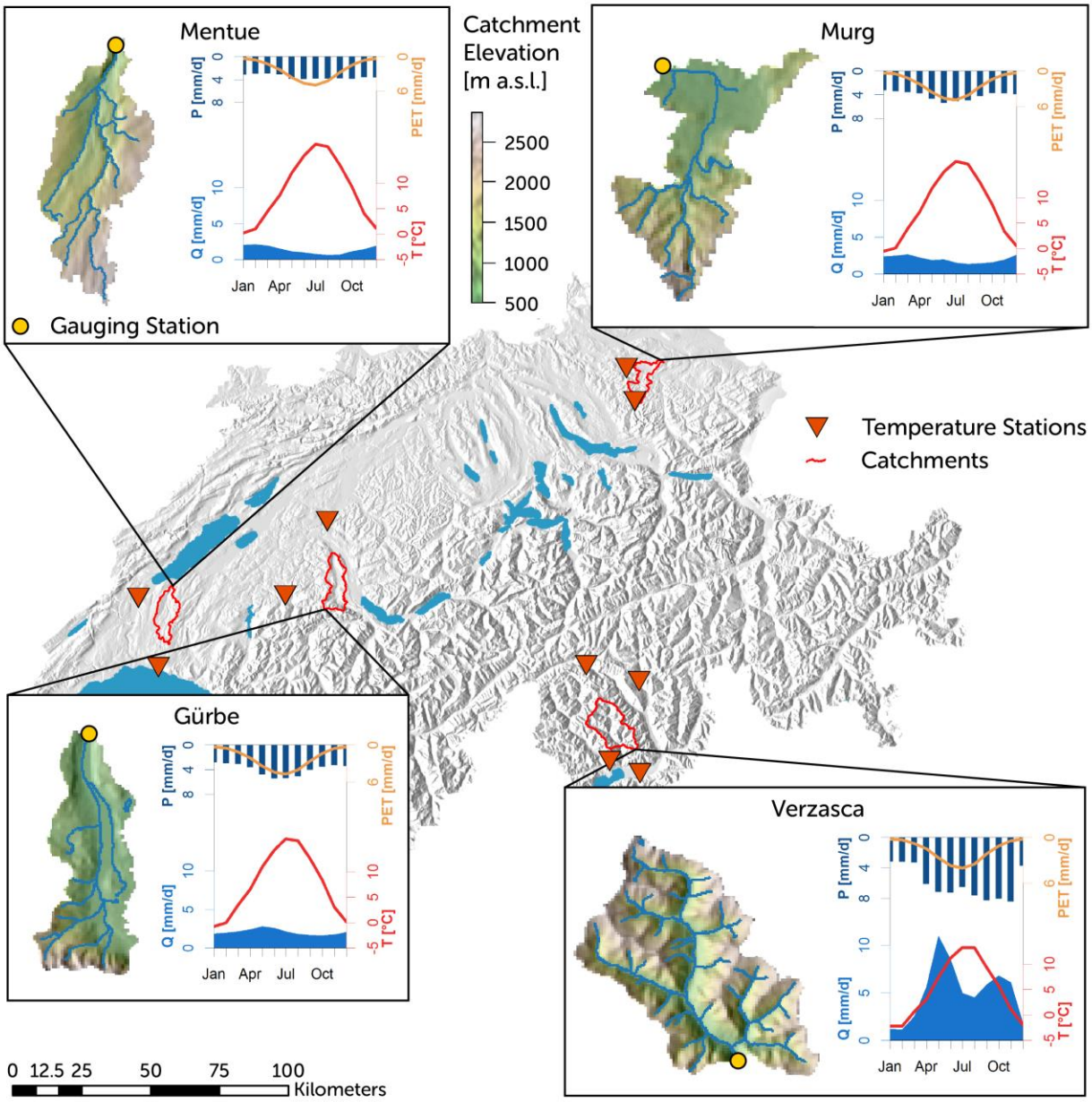


Figure 1 Map of Switzerland showing the location of the four catchments and the weather stations used to derive the temperature data. For each catchment, monthly average precipitation (P), streamflow (Q), temperature (T) and potential evapotranspiration (PET) are shown for the period 1974-2014, except for Verzasca for which the period 1990-2014 was used.

2.3 Measured data

Water level and streamflow time series were obtained from the Swiss Federal Office for the Environment (FOEN). The 10-minute measurements were averaged to obtain hourly water

level and streamflow time series. Hourly areal precipitation sums were obtained from the CombiPrecip dataset of MeteoSwiss (Sideris et al., 2014). The data for the years 2011 and 2013 suggest an unrealistic high runoff-rainfall ratio (>0.9) for the Verzasca catchment and were, thus, excluded from all simulations. A possible reason is that the weather stations are located outside the catchment and that precipitation is highly variable in this alpine terrain. Furthermore, the station data used in the CombiPrecip dataset are not corrected for wind undercatch, which can lead to errors of up to 40 % in winter for windy locations in Switzerland (Sevruk, 1985).

The hourly temperature at the mean elevation of the catchment was calculated from data from nearby weather stations (see Table 1 and Figure 1) using Thiessen polygons and a lapse rate of -6 °C per 1000 m. The potential evapotranspiration was calculated using the day of the year, the latitude and the temperature following the approach of McGuinness & Bordne (1972). We chose this simple model because more physically based potential evapotranspiration models would require more input data, which are not available with a satisfying spatial resolution in alpine terrain.

2.4 Selection of years for model calibration and validation

To obtain information on the influence of wetness conditions on the value of citizen science derived WL-class data for model calibration, we selected for each catchment an average, a dry and a wet year for model calibration and validation. For the average year, we selected the two years within the 2006-2014 period (the period with available hourly precipitation data at the time of the study) for which the total summer streamflow (July-September) was closest to the average summer streamflow for the 1974-2014 period. For the wet and the dry year, we selected the two years with the highest and lowest streamflow sum during the summer, respectively. For the calibration, we used the years that were second closest to the average, highest, or lowest value; for the validation, we used the year that were closest to the average and the years with the highest and lowest total streamflow during the summer (Table 1). Even though citizen science projects can obtain long-term data (e.g., the Audubon Christmas Bird Count has collected data for more than 100 years; Meehan et al., 2019), we wanted to test the value of one year of citizen science derived WL-class data for hydrological modelling because in reality most studies do not have time to obtain more extended time series.

2.5 Synthetic data

2.5.1 WL-class time series

We assume that the water level class observations are made at the catchment outlet. In order to determine the effect of the number of classes, we split the water level records from the FOEN into 2 to 10, 15 and 20 classes, resulting in 11 different WL-class time series per catchment. The WL-classes could, for instance, be obtained from a photograph of the stream with a sticker of a staff gauge added to it. The case with ten classes corresponds to the "virtual staff gauge" approach used in the CrowdWater app (Seibert et al., 2019; see example in Figure 2). The class borders were set at equal water level intervals between the 5^{th} and 95^{th} percentile of the water level record for the period for which the rating curve did not change and included the calibration years (Table 2). The cumulative frequency distribution of the water levels was approximately linear between the 5^{th} and 95^{th} percentile for all four catchments. Water levels below the 5^{th} and above the 95^{th} percentile would likely be below or above the virtual staff gauges set by the citizen scientists and were assigned to the lowest and highest WL-classes, respectively (Figure 2 and *Figure 3*).

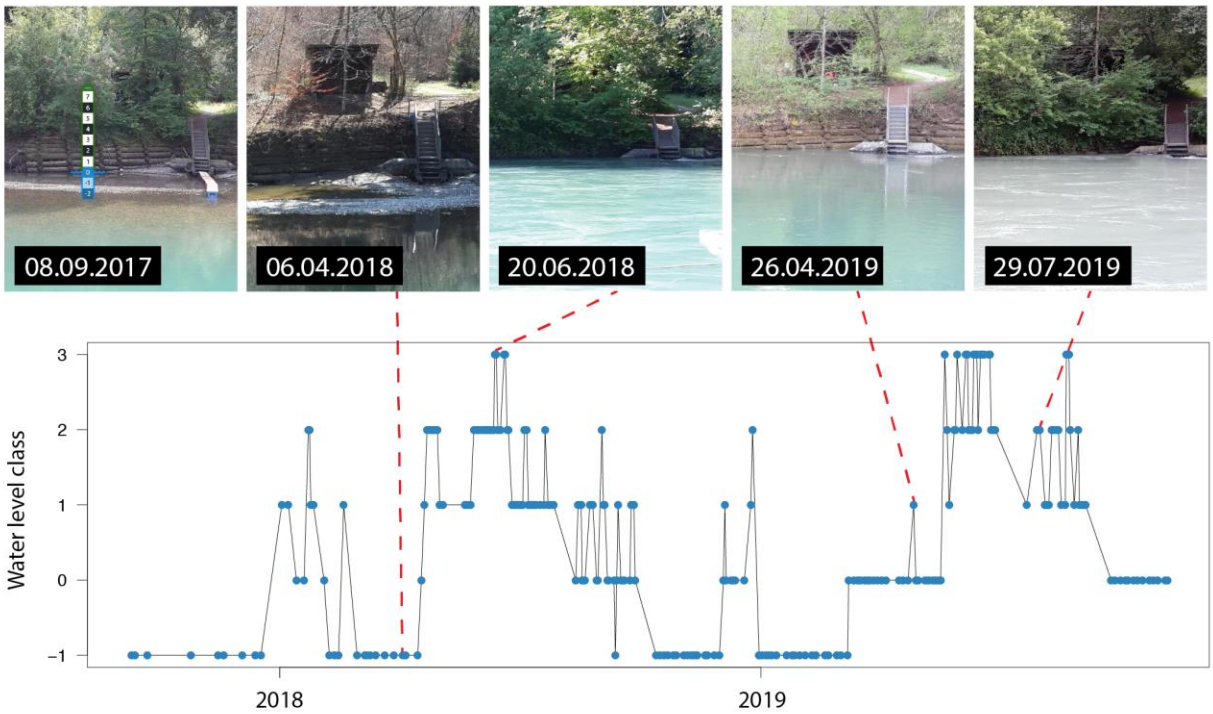


Figure 2 Timeseries of WL-Classes at the Aare river in Zollikofen, Switzerland based on the virtual staff gauge inserted on the reference picture (left picture in the upper row of the figure), which can then be used to estimate the water level class at the later dates (other pictures in the upper row). The entire time series of observations for this location can be found at: https://www.spotteron.com/crowdwater/spots/141766. Note that this time series illustrates the water level class data that can be observed by citizen scientists; we did not use this time series in the modelling described in this study. All photos were taken by Auria Buchs.

Table 2. The time periods of the water level records that were used to determine the WL-class boundaries and the dry, average and wet years chosen for model calibration and validation. The rating curves did not change considerably during the selected time period to determine the WL-class boundaries.

	Murg	**Guerbe**	**Mentue**	**Verzasca**
Period used for class definition	1974-2014	1996-2009*	1974-2014	1990-2013
Calibration years				
Dry	2013	2011*	2010	2013
Average	2008	2008	2006	2007
Wet	2007	2007	2014	2011
Validation years				
Dry	2009	2013	2009	2010
Average	2011	2006	2013	2006
Wet	2014	2014	2007	2008

* *For the Guerbe catchment, the dry calibration (2011) year occurred in a period after the rating curve changed so that there was a systematic shift in the water level data. Therefore, the class borders were determined for this period separately. For the validation period, we used streamflow data that were calculated with an adapted rating curve and therefore did not include this shift.*

2.5.2 Observation scenarios

We created water level and WL-class time series for observation scenarios that differed in the number of observations and the clustering of the observations throughout the year (Table 3). We used the same observation scenarios as Etter et al. (2018) for comparability For the *Crowd52* and *Crowd12* scenarios, we assigned higher probabilities to periods when people are more likely to be outdoors (i.e., a higher probability for summer than winter, a higher probability for weekends than weekdays, and a higher probability outside office hours; see Table 3 in Etter et al., 2018). This led to a larger number of observations during the summer for the *Crowd52* scenario than the *Weekly* scenario (median of 33 observations between May and September for *Crowd52* vs 22 for *Weekly*) and for *Crowd12* vs. the *Monthly* data (median of 8 for *Crowd12* vs. 5 for *Monthly*). In citizen science projects, the number of contributions will vary but based on our experience in the CrowdWater project, we assume that these scenarios cover a wide range of plausible cases.

In addition to the scenarios of Etter et al. (2018), we added the daily resolution for comparability with the results of van Meerveld et al. (2017). Daily data are not likely for citizen science

projects but near-daily data are possible: in CrowdHydrology 347 observations per year were made in the location with most contributions (Lowry et al., 2019). The location with most contributions in CrowdWater receives on average one observation every 1.2 days and in the spot in Figure 2 there was on average one observation every 3.2 days). The hourly water level data represent data from a water level logger, while hourly WL-class data could potentially be obtained from webcam images.

Table 3 The different scenarios for the temporal resolution of the observations used in this study, with the number of data points in one year of data (n)

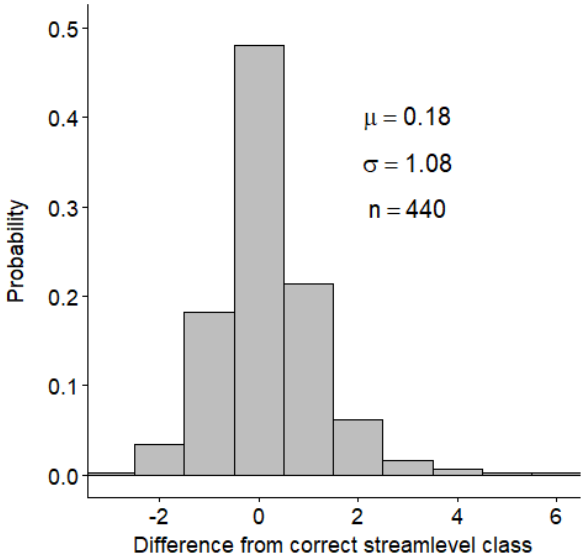
Hourly	One data point per hour ($8760 \leq n \leq 8784$, depending on the year)
Daily	One data point every day ($365 \leq n \leq 366$), randomly between 6 am and 8 pm
Weekly	One data point per week, every Saturday, randomly between 6 am and 8 pm ($52 \leq n \leq 53$)
Monthly	One data point per month on the 15th of the month, randomly between 6 am and 8 pm (n=12)
IntenseSummer	One data point every other day between July and September, randomly between 6 am and 8 pm (~15 observations per month, n = 46).
WeekendSummer	One data point each Saturday and each Sunday between May and October, randomly between 6 am and 8 pm ($52 \leq n \leq 54$)
WeekendSpring	One data point on each Saturday and each Sunday between March and August, randomly between 6 am and 8 pm ($52 \leq n \leq 54$)
Crowd52	52 data points (in order to be comparable to the *Weekly*, *IntenseSummer*, and *WeekendSpring* time series), between 6 am and 8 pm
Crowd12	12 data points (comparable to the *Monthly* data), between 6 am and 8 pm

2.5.3 Adding errors to the WL-class time series with 10 classes

Citizen-science-derived data likely contain errors. We assessed the typical errors in WL-class observations in a series of field surveys (Strobl et al., 2019a). We analysed 440 estimates of WL-classes from citizens who compared the water level in the stream that they were looking at to a photo of the same stream taken at an earlier time with a sticker of a staff gauge with ten classes added to it (the first photo in Figure 2 shows an example). Nearly half (48 %) of the participants chose the right class (as determined by experts) and 40 % were off by only one class (Strobl et al., 2019a). The errors (i.e., the difference between the reported WL-class and the

actual WL-class as determined by experts) were approximately normally distributed (Figure 3). We used these discrete class value probabilities to add random errors to each WL-class data point for the scenarios with 10 WL-classes. The same probability of errors was used for all four watersheds and years. In addition to this error, hereafter referred to as large error, we also created two time series with reduced errors to consider possible benefits of training or error-filtering (e.g. via reassessment of the WL-class data by multiple volunteers based on a comparison of images; Strobl et al., 2019b):

- Large error: typical errors of citizen scientists, i.e., random errors according to the normal distribution of errors from the survey of Strobl et al. (2019), as shown in Figure 3.
- Medium error: random errors according to the normal distribution with the standard deviation divided by two.
- Small error: random errors according to the normal distribution with the standard deviation divided by four.
- No error: The ten classes based on water level measurements by the FOEN, which are considered to be error-free and the benchmark in terms of quality for WL-class data.



296

297 *Figure 3 Distribution of the errors in the WL-class estimates (i.e., the difference between the*
298 *reported WL-class and the actual WL-class, as determined by experts) from field surveys for nine*
299 *different locations. The data was obtained from* (Strobl et al., 2019a). *This distribution was used*
300 *to create WL-class time series with large errors.*

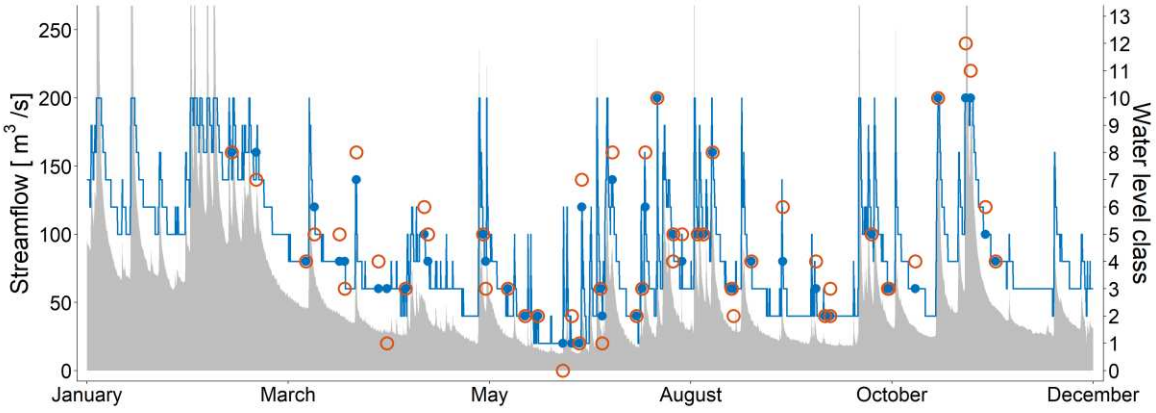


Figure 4 Observed streamflow at Mentue in 2014 (grey area), the hourly WL-class time series with 10 classes (blue line) derived from continuous water level data, and the synthetic data series for the Crowd52 scenario without any errors (blue dots) and large errors (orange circles) that were used for model calibration. The error distribution and formula used to add errors to the WL-Classes derived from the water level data are given in Figure 3.

301 ## 2.6 Model calibration

302 We calibrated the hydrological model for each of the synthetic data series (nine different
303 temporal resolutions, three error magnitudes with ten classes, and eleven class sizes without
304 errors) for each of the three calibration years for each of the four catchments. We also calibrated
305 the model for the nine different temporal resolutions of the water level data and the hourly
306 streamflow data for each year and catchment. For the calibration with measured streamflow we
307 used the overall performance index (P_{OA}; Finger et al., 2011). The P_{OA} is the mean of the Nash-
308 Sutcliffe efficiency for the streamflow (Nash and Sutcliffe, 1970), the Nash-Sutcliffe efficiency
309 for the log-transformed streamflow, the mean absolute relative error, and the volume error. For
310 each calibration with water levels or WL-classes, we optimized the Spearman rank correlation
311 coefficient (Spearman, 1904) for the relation between the synthetic WL-class data and the
312 simulated streamflow using a genetic optimisation algorithm (Seibert, 2000). The calibration-
313 ranges for the 16 parameters were based on their typical range and are the same as in Etter et al.
314 (2018). For each calibration, we used the preceding year as the warm-up period and calibrated
315 the model 100 times to account for parameter uncertainty. Each model calibration consisted of
316 3500 model runs and 1000 runs for local optimisation. This resulted in 100 parameter sets for
317 each of the three hourly streamflow calibrations (dry, average and wet year respectively), each of
318 the 27 water level simulations (3 years and 9 temporal resolutions), and each of the 378 WL-
319 class simulations (3 years, 9 temporal scenarios, and 3 error magnitudes plus 11 different class
320 sizes) per catchment, except for Verzasca for which only the average year was used for
321 calibration (Table 1). For the *Crowd52* and *Crowd12* datasets different realisations of the
322 observation times are possible and we, thus, randomly selected different observation times for
323 each of the 100 calibration trials. For these cases, the spread of the results is, thus, a combination
324 of parameter uncertainty and observation timing.

325 The Spearman rank coefficient cannot be computed if the WL-class dataset contained
326 data for only one class (i.e. due to a low variation in the water level data). This occurred for less
327 than 1% of all the scenarios studied here. For computation of the Spearman rank coefficient also
328 for these scenarios, the WL-class for the observation at the time of the highest streamflow was
329 manually changed to the next (higher) class.

330

2.7 Model validation

For each scenario, we used the 100 calibrated parameter sets to simulate the streamflow for the validation years. The validation performance was assessed using the overall performance index P_{OA}, as was done for the assessment of the value of uncertain streamflow data by Etter et al. (2018). We determined the median of the 100 P_{OA} values for each scenario and compared it to the median P_{OA} of the validation for the model calibrated with the observed streamflow data, which was considered the best possible model performance and thus the upper benchmark.

We similarly compared the median model validation performance for the different WL-class scenarios to the median validation performance of the model calibrated with the hourly water level time series. For each WL-class scenario, we also compared the validation performance to the validation performance of the model calibrated with water level data with the same temporal resolution in order to compare the value of citizen science based WL-class data and citizen science based water level data for model calibration. We used the one-sided paired Wilcoxon test to determine if the median model validation performance for the calibration with WL-class data was significantly worse than the validation performance for the model calibrate with the measured water level data. If there is not significant difference, then more easily scalable methods that do not require the installation of sensors, such as virtual staff gauges, are equally useful for model calibration as physical staff gauges. If the performance is significantly worse, it might be useful to invest in the installation of an actual staff gauge and have citizens report the water level from this staff gauge (Table 4).

The lower benchmark was defined as a situation were no streamflow, water level or WL-class data is available for model calibration. In wet environments, random parameters can result in surprisingly good model performance as long as the model reproduces the water balance. Therefore, the lower benchmark serves as the minimum model performance that can be expected based on the water balance alone (Seibert et al., 2018). Thus for the lower benchmark, we used the median performance of 1000 generated streamflow time series generated from the precipitation and temperature data in the validation period based on 1000 parameter sets that were selected randomly from the parameter ranges. We then compared the median validation performance of the models calibrated with streamflow, water level or WL-class data to the median model validation performance for the 1000 random parameter sets. We tested whether

361 the median model validation performance of the WL-class scenarios (for all nine calibration and
362 validation year combinations for all four catchments) was significantly better than the median
363 validation performance for the random parameters using the one-sided paired Wilcoxon test. We
364 considered the dataset useful for calibration when the median validation performance was
365 significantly better than for the random parameters (Table 4).

366 To determine the significance of differences in the median validation performance for the
367 different observation scenarios (i,e., different temporal resolutions) but the same number of WL-
368 classes and error category, we used a Kruskal-Wallis test with the Dunn Bonferroni post hoc test
369 with adjusted p-values for multiple comparisons (Bonferroni, 1936; Dunn, 1959).

370
371 *Table 4 Overview of the different model validation comparisons used to evaluate the value of*
372 *crowdsourced WL-class data. For each comparison the median validation performances were*
373 *compared using the one-sided paired Wilcoxon test. Significant differences are indicated by*
374 *filled squares in Figure 5 andFigure 6.*

Validation performance for calibration using WL-class data vs	Statistically significant difference in median P_{OA} value indicates:
Hourly streamflow data (upper benchmark)	A gauging station is more useful for model calibration than citizen science derived WL-class data using a virtual staff gauge
Hourly water level data	Installation of a water level recorder is more useful for model calibration than a virtual staff gauge that citizen scientists can use to determine the WL-class
Water level scenarios	Installation of a staff gauge from which citizens can read water levels is more useful for model calibration than a virtual staff gauge to determine the WL-class
Random parameter sets (lower benchmark)	Citizen science derived WL-class data have added value for model calibration

375

3 Results

3.1 Model performance for calibration based on hourly data

In general, the HBV model was able to reproduce the observed streamflow reasonably well when it was calibrated using the hourly streamflow data (upper benchmark). The median P_{OA} value for these calibrations was 0.82 (range: 0.66-0.88, with the lowest value for the calibration of the Guerbe for a dry year). The simulations for the validation period were not as good with a median P_{OA} of 0.64 (range: 0.19–0.83). The lowest validation P_{OA} value was for the Guerbe catchment when it was calibrated for the dry year and validated for the wet year (Table 5). These years had very different runoff-ratios (0.37 for the dry calibration year and 0.81 for the wet validation year; Table 1). The median validation performance (for all combinations of calibration and validation years) was also worst for the Guerbe catchment (P_{OA} = 0.55, range of other catchments 0.64 to 0.80; Table 5).

Table 5 Median validation performance (i.e., median P_{OA} values for the 100 parameters) for the different calibration and validation years when the model was calibrated with hourly streamflow data (upper benchmark).

Validation	**Dry**			**Average**			**Wet**			**Median**
Calibration	**Dry**	**Average**	**Wet**	**Dry**	**Average**	**Wet**	**Dry**	**Average**	**Wet**	**(of all year combinations)**
Murg	0.71	0.76	0.74	0.58	0.59	0.56	0.79	0.78	0.80	0.74
Guerbe	0.35	0.51	0.57	0.63	0.75	0.77	0.19	0.36	0.55	0.55
Mentue	0.40	0.41	0.23	0.64	0.64	0.75	0.66	0.65	0.73	0.64
Verzasca	0.63*	0.83	0.48*	0.52*	0.78	0.47*	0.65*	0.80	0.68*	0.80

* *These years had a runoff rainfall ratio >0.9 (see Table 1) and were not included in any of the other results.*

The median validation result of all model simulations based on model calibration using hourly water level data (median: 0.52; range: -0.39 to 0.78) was significantly worse than for the calibration with the hourly streamflow data ($p < 0.001$; Figure 5). The use of hourly water level data for model calibration caused the most noticeable decline in the median model validation performance for the Guerbe (P_{OA} relative to the upper benchmark: 0.45, range for the other catchments 0.75-0.92).

Calibration based on the hourly WL-class data led to a significantly worse median validation performance than calibration using streamflow data, regardless of the number of WL-classes (2-20 classes, all $p<0.001$). However, for the case without errors, the performance of the model calibrated with hourly WL-class data was not significantly worse than when the hourly water level data (i.e., hourly) were used for calibration, except for the case with ten classes due to outliers (see white squares in the second row on the top of Figure 5).

3.2 Effect of the number of observations and WL-classes (no-error case)

In general, the model validation performance was poorer when the model was calibrated with fewer water level or WL-class observations. Overall, the dataset with 52 crowdsourcing-like observations (*Crowd52*) led to the best model validation performance of all datasets with on average one observation per week. The scenario with two observations each weekend between March and August (*WeekendSpring*) and the scenario with regularly spaced weekly observations (*Weekly*) led to the next best model performance. Although the median validation performance for the models calibrated with the *WeekendSpring* data was always higher than for the model calibrated with observations each weekend from May to October (*WeekendSummer*), or every other day from July to September (*IntenseSummer*) (Figure 4), this difference was not statistically significant according to the Dunn-Bonferroni test (adjusted p-values were all >1.0).

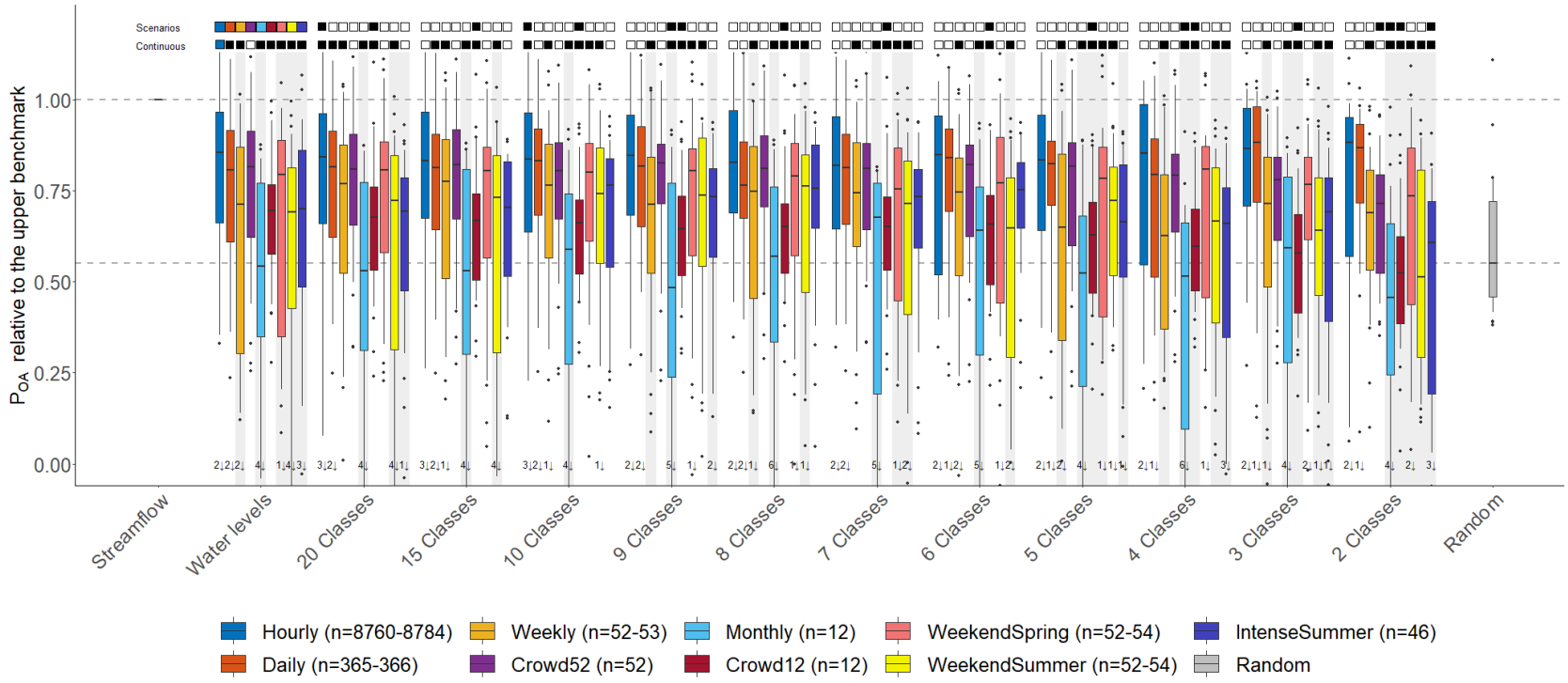
As one would expect, the model validation performance decreased slightly when the number of WL-classes decreased, but the effect depended on the temporal resolution of the data used for calibration (Figure 5). For two water level classes, only the scenarios *Hourly*, *Daily* and *Crowd52* led to similar model validation performances as the continuous water level data. For all other scenarios, performances were significantly worse ($p \leq 0.03$).

When daily WL-class data were used, the model validation performance was only for the cases with 15 and 20 classes significantly worse than the performance of the model calibrated with continuous water level data (p-values = 0.03 and 0.02, Figure 5). This was largely due to two outliers in both cases in the Guerbe catchment with P_{OA}-values between -0.18 and -0.4 or scores relative to the P_{OA} of the upper benchmark between -0.5 and -1.9. The validation performance of the model calibrated with the temporally discontinuous *Crowd52* water level or

WL-class datasets was never significantly worse from the validation performance of the model calibrated with the continuous water level data. The validation performance of the scenario focused on summer (*IntenseSummer*) was not significantly worse than the validation performance of continuous water level data if five or more WL-classes were used. The median model validation performance for the scenario with two observations each weekend between March and August (*WeekendSpring*) with 3, 4, 6, 15, and 20 WL-classes was not significantly different (p-values: 0.05-0.08) to the performance of the model calibrated with the hourly water level data either. This was also the case for the observations every other day between July and September (*IntenseSummer*) with at least 5 WL-classes (p-values: 0.06-0.27). For all the other scenarios the model validation performance was significantly worse than for the calibration with continuous water level data (see black squares in the second row above the main plot in Figure 5).

Calibration with discontinuous WL-class data led in only very few cases to a significantly poorer model performance than calibration with temporally discontinuous but precisely measured water level data: the *Crowd12* scenario regardless of the number of WL-classes, the *Monthly* scenario with 2, 4, and 9 classes, and the *Crowd52*, *WeekendSpring*, *WeekendSummer* and *IntenseSummer* scenario with two classes (see black squares in first row above Figure 5).

The validation performance for the model calibrated with the *Hourly*, *Daily* and *Crowd52* datasets was always better than the lower benchmark. However, monthly WL-class data (*Monthly*) never improved the validation performance compared to the lower benchmark (Figure 5). For five or fewer classes, there were more scenarios for which the model did not perform significantly better than the lower benchmark, e.g., the *Weekly*, *Crowd12*, *WeekendSpring*, *WeekendSummer* and *IntenseSummer* scenarios. However, the p-values were close to 0.05 and therefore the significance test results differed for the different number of WL-classes. The model performance for the *IntenseSummer* and *WeekendSpring* scenarios did not systematically improve with an increasing number of classes, hence model performance for these scenarios was best when eight or nine WL-classes were used.



457

458 *Figure 5 Box plots of the validation performance of the HBV-model calibrated with synthetic WL-class data (different temporal*
459 *resolutions and different numbers of WL-classes) relative to the performance of the model calibrated with hourly streamflow data. The*
460 *lower benchmark (in grey) represents the median performance of the model run with 1000 randomly selected parameter sets. The grey*
461 *background shading highlights the scenarios for which the median model performance was not significantly better than for the lower*
462 *benchmark. The filled squares at the top of the graph indicate cases where the median validation performance for the model*
463 *calibrated with WL-class data was significantly worse compared to the calibration with water level data with the same temporal*
464 *resolution (top row) and compared to the calibration with continuous (hourly) water level data (second row); empty squares indicate*
465 *no statistically significant difference based on the one-sided paired Wilcoxon test. All scenarios led to a significantly worse model*
466 *validation performance than calibration with continuous streamflow data. The WL-classes were equally sized and assumed to be error*
467 *free. The box extends from the 25th to 75th percentile and the whiskers extend to the 10th and 90th percentile. The black line inside the*
468 *box represents the median. Numbers at the bottom indicate outliers with a relative P_{oa}<0.00.*

3.3 Effect of errors in WL-class estimates with ten classes

Including errors in the WL-class data resulted in only a minor decrease in the overall model validation performance. This effect was particularly small compared to the effect of the temporal resolution of the data used for model calibration (Figure 6). For all *Hourly*, *Daily*, *Crowd52*, *Crowd12*, and the *WeekendSpring* cases with ten WL-classes, the model validation performance was better than the lower benchmark, even when large errors were included in the calibration data (Figure 6). The effect of errors on model validation performance was most substantial for calibration with the *Weekly* and *IntenseSummer* datasets for which the scenarios with medium and large errors were not significantly better than the lower benchmark. The addition of medium or large errors also caused the validation performance for the model calibrated with the *IntenseSummer* data to become significantly worse than the model calibrated with continuous hourly water level data (Figure 6). The performance of the model calibrated with the *Daily* data also became only significantly worse than the model calibrated with continuous water level data when small or medium errors were included. The model validation performance for calibration with *Hourly* and *Crowd52* WL-class data was not significantly worse than the validation performance for calibration with continuous water levels, even with large errors (Figure 6).

The median validation performance of the model calibrated with the discontinuous WL-class data remained similar to the performance of the model calibrated with discontinuous water level data, except for *Crowd12* and *Hourly* WL-classes (again due to the large outliers in the Guerbe catchment) for which the calibration with WL-class data with errors led to a significantly worse validation performance than calibration with discontinuous water level data.

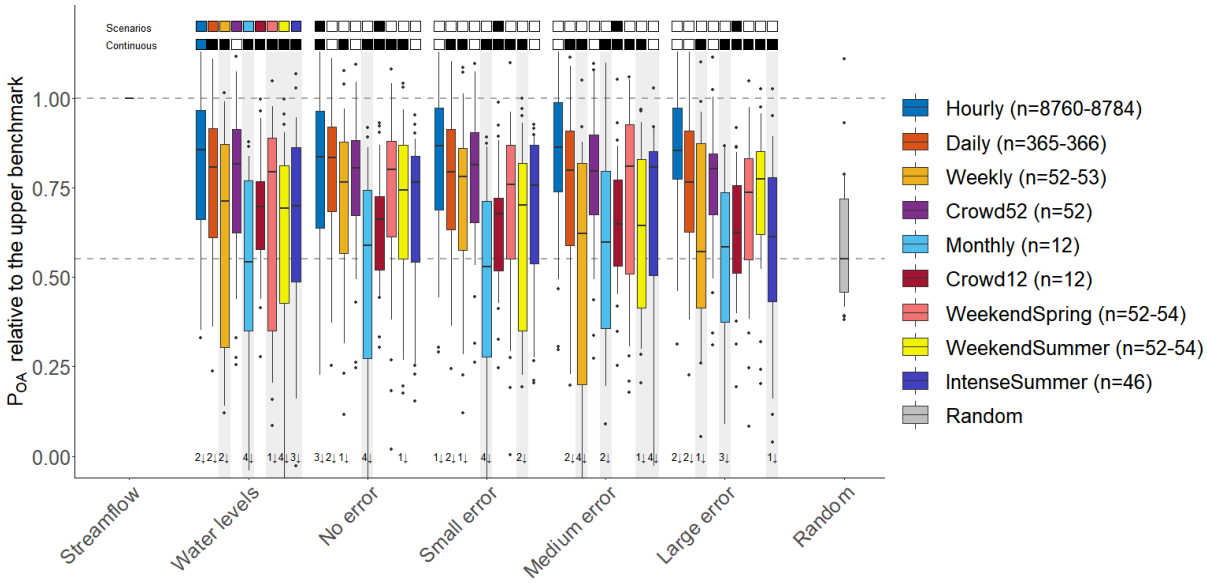


Figure 6 Box plots of the model validation performance of the HBV-model calibrated with water level data with different temporal resolutions and the synthetic WL-class data (ten equal sized classes) with different temporal resolutions and different errors, relative to the validation performance of the model calibrated with hourly streamflow data (upper benchmark). The lower benchmark shown (in grey) is the median validation performance of the model run with 1000 random parameters. The grey shading indicates a median model performance that is not significantly better than the lower benchmark ($p>0.05$). The filled black squares at the top of the graph indicate cases where the median validation performance for the calibration with WL-class data is significantly worse than the calibration with water level data with the same temporal resolution (top row) and compared to continuous water level data (second row); empty squares indicate no statistically significant difference based on the one-sided paired Wilcoxon test.

3.4 Effects of variability in WL-class data on model performance

For the *Crowd52* scenarios, there were 100 realisations for every catchment and year. This allowed us to explore the effect of the distribution of the WL-class observations on model performance. For the wet years with more streamflow in summer, there was a more balanced distribution of data points across the classes than for the dry years. For the wet years, 14% of all data points were in the lowest class, 19% in the second class, and 22% in the third class when ten WL-classes were used. The corresponding numbers were 20, 24 and 17% for the average year and 45, 16 and 14% for the dry years. For the *Crowd52* scenario with ten classes and no errors, the median validation performance for parameter sets obtained by calibration with data from the

513 wet year (median P_{OA} = 0.54) and the average year (median P_{OA} = 0.56) were significantly better
514 than for the calibration with data from the dry years (median P_{OA} = 0.44, both $p \leq 0.001$).

515 We also compared the model validation performance of *Crowd52* scenarios for WL-class
516 data based on 10 classes and without errors with a different number of observations in classes 1
517 and 2 (i.e. observations during baseflow conditions). If more than half of the observations were
518 in classes 1 or 2, model validation performance was significantly worse than if there were
519 relatively fewer observations for classes 1 and 2 (and thus more observations for classes 3-10).
520 This indicates that the model performance increases when there are more observations for the
521 higher WL-classes. However, for the *Crowd52* scenario, there was no correlation between the
522 variance in WL-classes used for calibration and model validation performance for the resulting
523 100 calibrated parameter sets (adjusted r-squares all ≤0.01). This is likely due to the large
524 variability in the individual parameter sets and their effect on model performance because for the
525 *Crowd52* scenario only one parameter set was obtained for each observation scenario.

526

527 **4 Discussion**

528 With this study, we extended our understanding of the value of uncertain data for
529 hydrological model calibration. The usefulness of WL-class data for model calibration was
530 shown earlier for continuous WL-class data for a large number of catchments in the US by van
531 Meerveld et al. (2017). Here we show that even discontinuous WL-class data are useful for
532 model calibration. We used the HBV model for the analyses but argue that the findings would be
533 similar for other bucket-type hydrological models. For physically-based spatially distributed
534 models that are used without calibration, WL-class data might still be useful for model
535 evaluation. The results are likely different for arid regions, where rivers only flow at certain
536 times of the year, as Seibert & Vis (2016) showed that model parameterizations based on
537 calibration against water level data were less suitable to simulate streamflow for arid regions
538 than for humid regions.

539

4.1 Value of WL-class data for hydrological model calibration

Usually hydrological models are calibrated using streamflow data derived from water level measurements and a rating curve. In practice, this is the most expensive method to obtain stream-related data but it also leads to the best validation results (which is why we consider this the upper benchmark). Continuous water level measurements are easier to obtain; water level loggers have become cheaper and can now easily store a year of data. However, the installation of water level loggers still requires some investment and maintenance, particularly in steep mountainous terrain where the stream channel may change frequently due to scour and deposition. The different temporal observation scenarios with precise water level data represent the case when a physical staff gauge is placed in a stream and passers-by read the level and transmit their observation, as it is for example done in the CrowdHydrology project (Lowry & Fienen, 2013), Cithyd (www.cithyd.com; Balbo & Galimberti, 2016) and a project in Kenya (Weeser et al., 2018).

The median validation performance for the models calibrated using WL-class data was worse than for the model that was calibrated using streamflow data but as good as using water level data with the same temporal resolution. Even for the realistic citizen science scenario *Crowd52*, the validation performance was not significantly worse than when hourly water level data that are recorded with a water level logger are used for calibration. These results suggest that while traditional streamflow measurements are most informative for hydrological model calibration and continuous hourly water level data certainly have their value, observations of WL-classes, e.g., based on virtual staff gauges (Seibert et al., 2019), are also valuable for model calibration and can lead to reasonable streamflow simulations when streamflow data are not available. Model calibration with WL-class data can be used to transform the measured WL-classes into streamflow time series and, thus, be used to derive useful information, such as hydrologic signatures (e.g., runoff-rainfall ratios). The use of regionalized parameter values (Andréassian et al., 2014) would be an alternative approach to derive streamflow estimates for ungauged basins. This approach, however, is also subject to uncertainties as the transfer functions will only be approximations (Hundecha et al., 2002). This was therefore not part of this study but it raises the interesting question whether a few water level class observations can improve regionalized parameter sets for areas where there are no other streamflow data.

570

571 4.2 Effects of timing of the WL-class observations and errors on model performance

572 The number of observations and the timing of the observations in the year had a larger
573 influence on model performance than errors in the WL-class observations. Errors generally had
574 the largest effect on model performance when few observations were available for calibration, as
575 was the case for the *Monthly* and the *Crowd12* scenarios (Figure 6). Compared to the effect of
576 errors in streamflow estimates on model validation performance (Etter et al., 2018), the effect of
577 errors in the WL-classes was minor. This can be explained by the fact that there are no extreme
578 outliers in the WL-class data and that the errors in the WL-class estimates are smaller than those
579 for streamflow estimates (Strobl et al., 2019a). Even for the large error case, 48 % of the
580 observations were in the correct class and 88 % of the observations were within ±1 class of the
581 correct class (Strobl et al., 2019; Figure 3).

582 Although there was a general trend of increasing model performance with an increasing
583 number of observations, the timing of the observations within the year also had a substantial
584 effect on model performance. The validation performance for the model calibrated with
585 *Crowd52* data (i.e., with more observations in summer) was comparable to the performance of
586 the model calibrated with hourly water level data, regardless of the number of classes. On the
587 other hand, the model validation performance of the model calibrated with *Weekly* data was
588 significantly worse than the performance of the model calibrated with hourly water level data,
589 even when using 20 water level classes. This is contrary to the results for uncertain streamflow
590 observations of Etter et al. (2018), where *Weekly* data resulted in a better model validation
591 performance than *Crowd52* data. For WL-class estimates, it is probably beneficial to obtain
592 observations that cover a larger variation in streamflow magnitudes than for streamflow directly
593 because it takes a relatively large change in the actual water level (and thus also streamflow) to
594 change one WL-class. More variations in streamflow occur more often during wet years with
595 higher flows, leading to the significantly better validation performance for the wet or average
596 years compared to the dry years for the *Crowd52* dataset. A denser sampling strategy during
597 summer is also more likely to catch more of the variation in streamflow, leading to the better
598 model validation performance for the model calibrated with *Crowd52* data compared to *Weekly*
599 data and for the *IntenseSummer* data (with observations every other day during June, July and

August) compared to *WeekendSummer* data. This also explains why the *IntenseSummer* scenario led to a similar performance as *WeekendSpring*, even though that scenario covers more streamflow variation during spring. The median model validation performance for the calibration with the *WeekendSpring* data was higher than for the *WeekendSummer* data and comparable (i.e., not significantly worse) to the validation performance of the model calibrated with the hourly water level data, while calibration with the *WeekendSummer* data led to a significantly worse model performance than when hourly water level data were used for calibration.

4.3 Influence of the number of WL-classes on model performance

The staff gauges in the survey of Strobl et al. (2019) had ten classes and, thus, the errors used in this study are typical for staff gauges with ten classes. However, in practice fewer classes will be used for many locations as the virtual staff gauges that are inserted in the pictures are often too large, so that it is unlikely that the water level will reach the highest classes (Seibert et al., 2019). Our results indicate, however, that even when the water level fluctuates in only two or three classes, such data can be informative for model calibration if there is on average at least one observation per week. The influence of errors on such observations might, however, be larger than when all ten classes are used (although the chances for large observation errors are likely smaller for very large virtual staff gauges).

The benefit of using more than four to five WL-classes (depending on the scenario) for model calibration was negligible. This is roughly in line with the findings of van Meerveld et al. (2017), who showed for continuous WL-class data for about 600 catchments in the US that there was hardly any improvement in model performance if more than five WL-classes were used. However, the observation scenario affects this result, e.g., for the *Weekly* scenario the results tended to be more stable when ten classes or more were used (Figure 5). However, the results of the scenarios with observations during summer (*WeekendSummer* and *IntenseSummer*) suggest that in terms of model performance it is not necessarily helpful to have more WL-classes. Especially in summer, when extended periods of low flows can be expected, eight to ten classes might provide the model enough degrees of freedom to perform well in a validation year that is different from the calibration year, whereas more WL-classes can lead to overfitting of the model to the calibration period. During periods of low flow, the water level will vary across more classes when more classes are used (and individual WL-classes are thus smaller), which might

lead to overfitting of the model for that particular year and result in calibrated parameter sets that do not perform well during other years.

4.4 Implications for citizen science projects

For citizen science projects, where the data quality often is an important issue (Show, 2015), clear and straightforward procedures help to ensure good data quality (Cohn, 2008). Based on the results of this study, a simple approach using a virtual staff gauge with ten classes (as implemented in the CrowdWater app; Seibert et al., 2019) can provide data that are useful for model calibration. The WL-class estimates seem to be superior for citizen science projects than streamflow estimates as indicated by the significantly better model performance of the *Crowd52* and *WeekendSpring* data sets compared to the calibration using random parameters, even when the errors in the observations were large. This was not the case for streamflow estimates, for which large errors hampered the usefulness of the data for model calibration (Etter et al., 2018).

The lack of an increase in model performance for most scenarios when more than four to five WL-classes were used indicates that the exact number of WL-classes does not significantly impact model calibration if at least four to five WL-classes are used. It also suggests that model performance should not be impacted dramatically if citizen scientists do not perfectly place the virtual staff gauge in the CrowdWater app so that the water level fluctuations do not cover all classes, as long as the water level fluctuates over at least four classes. In some cases, even fewer classes might be sufficient, especially if there is on average more than one observation per week, which is not unlikely when dedicated volunteers submit observations (Lowry et al., 2019).

More observations at higher flows and therefore higher WL-classes improved the model performance. Based on the significantly worse model performance for the *Crowd52* scenarios for which the percentage of observations during baseflow conditions was larger than 50 % compared to scenarios for which this was less, we conclude that it is beneficial to encourage observations during times with larger water level fluctuations. This finding was also supported by the better model performance for the model calibrated with the *Crowd52* data for wet or average years compared to dry years. This is also the case when physical staff gauges (instead of virtual staff gauges for WL-class observations) are used. For some streams, particularly those with a flashy response, it may be difficult to get observations at high flow conditions because people are less likely to be outside and willing to stop to submit an observation. For other streams, this is

possible, particularly when dedicated volunteers contribute regular observations because the high water levels are also very interesting for them (see example in Figure 3). A larger number of observations during these high flow periods can be obtained by sending push-messages if there is an app, text messages, e-mails or social media posts.

Although the differences in model validation performance for the discontinuous water level and WL-class data were in most cases not statistically significant (Figure 5 Figure 6), there are advantages and disadvantages for both methods. The advantage of a real staff gauge is that citizens who pass by a location of interest may notice the staff gauge and are more directly invited to participate in the project. With the virtual staff gauge approach, this is not the case for people who have not installed the app yet (or haven't looked at the map of existing observation sites). Signpost to encourage participation could be used to highlight the virtual staff gauge site but this requires additional effort (for the project administrators to install the sign and for the citizen scientists to first download the app). Another advantage of a physical staff gauge is that at locations where the streambed doesn't change, the water level observations could be transformed to streamflow once a rating curve is available for that location. This is also possible for the WL-class data but of course results in an upper and lower bound of the streamflow for each WL-class observation (Strobl et al., 2019a). When the riverbed changes drastically either a new (virtual) staff gauge needs to be 'installed' or the timeseries need to be considered separately. In case the data are used for model calibration, the alternative (though less preferable option) might be to use different parameter sets for the different periods.

The advantage of the virtual staff gauge approach is that it is easily scalable because only the citizen scientist needs to be at the location to set up a station and no equipment, permission and local maintenance are required (Seibert et al., 2019). Of course, the use of an app, text messages or even paper forms and mailboxes, can also be combined. From a data quality perspective, the advantage of a virtual staff gauge approach to collect WL-class data using an app (e.g., the CrowdWater app) is that observations can be stored offline if no cellular network connection is available and can be uploaded later. Furthermore, the observations come with a picture of the situation, which allows some form of checking the data quality and allows further analysis using image recognition techniques. This is also possible for water level observations at real staff gauges but when only text messages are sent, the recipient must trust the sender that the water level reading and the time of the observation are correct.

5 Conclusions

We studied the potential value of water level (WL)-class data that can be collected in citizen science projects for hydrological model calibration. Such data will be irregular in time, affected by errors and less precise than water level data. Our findings show that citizen science approaches to collect water level data using virtual or physical staff gauges with a few classes or precise markings are a promising way to obtain useful data for hydrological modelling in data-scarce catchments.

The results from the synthetic datasets indicated that time series with on average one WL-class observation per week over a one-year period provides valuable information for calibration of a lumped bucket type model if there are four or more classes. Typical errors in the WL-class estimates for citizen science projects (Strobl et al., 2019) did not impact model performance considerably. Although the validation performance of the model calibrated with synthetic WL-class data with realistic frequencies for citizen science projects was not as good as when streamflow data were used for calibration, the performance was comparable to calibration with data collected with water level loggers or physical staff gauges with precise markings. The WL-class observation approach has the advantage of being easier to implementation and more scalable because it does not require any physical installations (and, thus, no special equipment, permits or maintenance).

6 Acknowledgements

We thank all citizens who participated in the surveys to determine the typical error in the WL-class observations, the Swiss Federal Office for the Environment for providing the streamflow and water level data, MeteoSwiss for the provision of the weather data, Maria Staudinger for compiling the HBV input files, Marc Vis for HBV-model support, and the UZH Science Cloud for the provision of the infrastructure for cloud computing. The CrowdWater project is funded by the Swiss National Science Foundation (project no. 163008).

7 Shared data

The streamflow and water level data used for this study were obtained from the Swiss Federal Office for the Environment (FOEN); the station numbers are given in Table 1. The weather data were obtained from MeteoSwiss. The data repository for this study (Etter et al., 2019) contains the streamflow data from the FOEN for the selected calibration and validation years, the water level data, and WL-class data for all error types and observation scenarios, the model input and output files, the table with the parameter ranges, and the R-scripts.

8 References

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., & Anderson, S. E. (2017). The Accuracy of Citizen Science Data: A Quantitative Review. *The Bulletin of the Ecological Society of America*, *98*(4), 278–290. https://doi.org/10.1002/bes2.1336

Andréassian, V., Bourgin, F., Oudin, L., Mathevet, T., Perrin, C., Lerat, J., et al. (2014). Seeking genericity in the selection of parameter sets: Impact on hydrological model efficiency. *Water Resources Research*, *50*(10), 8356–8366. https://doi.org/10.1002/2013WR014761

Aschwanden, H., & Weingartner, R. (1985). *Die Abflussregimes der Schweiz. Publikation Gewässerkunde* (Vol. 65). Bern, Switzerland: Geographisches Institut der Universität Bern, Abteilung Physikalische Geographie, Gewässerkunde, 1985.

Balbo, A., & Galimberti, G. (2016). Citizen hydrology in River Contracts for water management and people engagement at basin scale. In *COWM2016 - International Conference on Citizen Observatories for Water Management*. Venice.

Bergström, S. (1976). *Development and application of a conceptual runoff model for Scandinavian catchments*. (S. Bergström, Ed.), *SMHI Rapporter* (Vol. 52). Norrköping, Sweden: SMHI Norrköping, Report RH07.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*. Florence: Istituto

superiore di scienze economiche e commerciali.

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., et al. (2014). Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science*, *2*(26), 21. https://doi.org/10.3389/feart.2014.00026

Christophersen, N., Neal, C., & Hooper, R. P. (1993). Modelling the hydrochemistry of catchments: a challenge for the scientific method. *Journal of Hydrology*, *152*(1–4), 1–12. https://doi.org/10.1016/0022-1694(93)90138-Y

Cohn, J. P. (2008). Citizen Science: Can Volunteers Do Real Research? *BioScience*, *58*(3), 192–197. https://doi.org/10.1641/B580303

Davids, J. C., van de Giesen, N., & Rutten, M. (2017). Continuity vs. the Crowd—Tradeoffs Between Continuous and Intermittent Citizen Hydrology Streamflow Observations. *Environmental Management*, *60*(1), 12–29. https://doi.org/10.1007/s00267-017-0872-x

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., et al. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, *10*(6), 291–297. https://doi.org/10.1890/110236

Dunn, O. J. (1959). Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics*, *30*(1), 192–197. https://doi.org/10.1214/aoms/1177706374

Etter, S., Strobl, B., Seibert, J., & van Meerveld, I. (2018). Value of uncertain streamflow observations for hydrological modelling. *Hydrology and Earth System Sciences*, *22*, 5243–5257. https://doi.org/https://doi.org/10.5194/hess-22-5243-2018

Etter, S., Strobl, B., Seibert, J., & van Meerveld, I. (2019). Data and R-Scripts for: Value of crowd-based water level class observations for hydrological model calibration. https://doi.org/10.5281/zenodo.3371308

Fekete, B. M., Looser, U., Pietroniro, A., & Robarts, R. D. (2012). Rationale for Monitoring Discharge on the Ground. *Journal of Hydrometeorology*, *13*(6), 1977–1986.

https://doi.org/10.1175/JHM-D-11-0126.1

Hundecha, Y., Zehe, E., & Bárdossy, A. (2002). Regional parameter estimation from catchment properties prediction in ungauged basins. *Predictions in Ungauged Basins: PUB Kick-Off*, (December).

Kampf, S., Strobl, B., Hammond, J., Annenberg, A., Etter, S., Martin, C., et al. (2018). Testing the waters: Mobile apps for crowdsourced streamflow data. *Eos*, *99*. https://doi.org/10.1029/2018EO096355

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560. https://doi.org/10.1002/fee.1436

Kundzewicz, Z. W. (1997). Water resources for sustainable development. *Hydrological Sciences Journal*, *42*(4), 467–480. https://doi.org/10.1080/02626669709492047

Lanfranchi, V. ., Wrigley, S. N. . N., Ireson, N. ., Wehn, U. ., & Ciravegna, F. . (2014). Citizens' observatories for situation awareness in flooding. *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*, (May), 145–154.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1–4), 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3

Lowry, C. S., & Fienen, M. N. (2013). CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists. *Ground Water*, *51*(1), 151–156. https://doi.org/10.1111/j.1745-6584.2012.00956.x

Lowry, C. S., Fienen, M. N., Hall, D. M., Stepenuck, K. F., & Paul, J. D. (2019). Growing Pains of Crowdsourced Stream Stage Monitoring Using Mobile Phones: The Development of CrowdHydrology. *Frontiers in Earth Science*, *7*(128), 1–10. https://doi.org/10.3389/feart.2019.00128

McGuinness, J., & Bordne, E. (1972). *A comparison of lysimeter-derived potential evapotranspiration with computed values*. Washington, DC: Agricultural Research Service - United States Department of Agriculture.

McMillan, H. K., Westerberg, I. K., & Krueger, T. (2018). Hydrological data uncertainty and its implications. *Wiley Interdisciplinary Reviews: Water*, *5*(6), e1319. https://doi.org/10.1002/wat2.1319

Meehan, T. D., Michel, N. L., & Rue, H. (2019). Spatial modeling of Audubon Christmas Bird Counts reveals fine-scale patterns and drivers of relative abundance trends. *Ecosphere*, *10*(4), e02707. https://doi.org/10.1002/ecs2.2707

van Meerveld, H. J., Vis, M. J. P., & Seibert, J. (2017). Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences*, *21*(9), 4895–4905. https://doi.org/10.5194/hess-21-4895-2017

Mulligan, M. (2013). WaterWorld: a self-parameterising, physically based model for application in data-poor but problem-rich environments globally. *Hydrology Research*, *44*(5), 748. https://doi.org/10.2166/nh.2012.217

Njue, N., Stenfert Kroese, J., Gräf, J., Jacobs, S. R., Weeser, B., Breuer, L., & Rufino, M. C. (2019). Citizen science in hydrological monitoring and ecosystem services management: State of the art and future prospects. *Science of The Total Environment*, *693*, 133531. https://doi.org/10.1016/j.scitotenv.2019.07.337

Pool, S., Viviroli, D., & Seibert, J. (2019). Value of a Limited Number of Discharge Observations for Improving Regionalization: A Large-Sample Study Across the United States. *Water Resources Research*, *55*(1), 363–377. https://doi.org/10.1029/2018WR023855

Seibert, J. (2000). Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, *4*(2), 215–224. https://doi.org/10.5194/hess-4-215-2000

Seibert, J., & Vis, M. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, *16*(9), 3315–3325.

https://doi.org/10.5194/hess-16-3315-2012

Seibert, J., & Vis, M. J. P. (2016). How informative are stream level observations in different geographic regions? *Hydrological Processes*, *30*(14), 2498–2508. https://doi.org/10.1002/hyp.10887

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125. https://doi.org/10.1002/hyp.11476

Seibert, J., & McDonnell, J. J. (2015). Gauging the Ungauged Basin: Relative Value of Soft and Hard Data. *Journal of Hydrologic Engineering*, *20*(1), A4014004-1–6. https://doi.org/10.1061/(ASCE)HE.1943-5584.000086

Seibert, J., Strobl, B., Etter, S., Hummer, P., & van Meerveld, (H. J.) Ilja. (2019). Virtual Staff Gauges for Crowd-Based Stream Level Observations. *Frontiers in Earth Science*, *7*. https://doi.org/10.3389/feart.2019.00070

Seibert, J., van Meerveld, H. J., Etter, S., Strobl, B., Assendelft, R., & Hummer, P. (2019). Wasserdaten sammeln mit dem Smartphone – Wie können Menschen messen, was hydrologische Modelle brauchen? *Hydrologie Und Wasserbewirtschaftung*, *63*(2). https://doi.org/10.5675/HyWa_2019.2_1

Sevruk, B. (1985). Systematischer Niederschlagsmessfehler in der Schweiz. In Sevruk B. (Ed.), *Der Niederschlag in der Schweiz* (pp. 31, 65–75). Beiträge zur Geologie der Schweiz - Hydrologie.

Show, H. (2015). Rise of the citizen scientist. *Nature*, *524*(7565), 265–265. https://doi.org/10.1038/524265a

Sideris, I. V., Gabella, M., Erdin, R., & Germann, U. (2014). Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. *Quarterly Journal of the Royal Meteorological Society*, *140*(680), 1097–1111. https://doi.org/10.1002/qj.2188

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Strobl, B., Etter, S., van Meerveld, I., & Seibert, J. (2019a). Accuracy of crowdsourced streamflow and stream level class estimates. *Hydrological Sciences Journal*, 1–19. https://doi.org/10.1080/02626667.2019.1578966

Strobl, B., Etter, S., van Meerveld, I., & Seibert, J. (2019b). The CrowdWater game: A playful way to improve the accuracy of crowdsourced water level class data. *PLOS ONE*, *14*(9), e0222579. https://doi.org/10.1371/journal.pone.0222579

Sy, B., Frischknecht, C., Dao, H., Consuegra, D., & Giuliani, G. (2018). Flood hazard assessment and the role of citizen science. *Journal of Flood Risk Management*, e12519. https://doi.org/10.1111/jfr3.12519

Weeser, B., Stenfert Kroese, J., Jacobs, S. R., Njue, N., Kemboi, Z., Ran, A., et al. (2018). Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. *Science of The Total Environment*, *631–632*, 1590–1599. https://doi.org/10.1016/j.scitotenv.2018.03.130